

Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections

Rubén G. Mateo^{1,2*}, Thomas B. Croat³, Ángel M. Felicísimo⁴ and Jesús Muñoz¹

¹Real Jardín Botánico (CSIC), Plaza de Murillo 2, E-28014 Madrid, Spain,

²Universidad de Castilla-La Mancha, Av. Carlos III s/n, Toledo, 45071, Spain, ³Missouri Botanical Garden, P.O. Box 299, St. Louis, MO 63166-0299, USA, ⁴Escuela Politécnica, Universidad de Extremadura, E-10071 Cáceres, Spain

ABSTRACT

Aim The presence-only data stored in natural history collections is the most important source of information available regarding the distribution of organisms. These data and profile techniques can be used to generate species distribution models (SDMs), but pseudo-absences must be generated to use group discriminative techniques. In this study, we evaluated whether the SDMs generated with pseudo-absences are reliable and also if there are differences in the results obtained with profile and group discriminative techniques.

Location Ecuador, South America.

Methods The SDMs were generated with a training data set for each of the five species of *Anthurium* and six different methods: two profile techniques (BIOCLIM and Gower's distance index), three group discriminative techniques [logistic multiple regression (LMR), multivariate adaptive regression splines (MARS) and MAXENT] and a mixed modelling approach genetic algorithm for rule-set production (GARP), which employs a combination of profile and group discriminative techniques and generates its own pseudo-absences. For LMR, MARS and MAXENT, three types of absences were generated: (1) random pseudo-absences in equal number to presences and excluding a buffer area around presences (except for MAXENT, which assumes that this background sample includes presences), (2) a large number (10,000) of random pseudo-absences, also excluding a buffer area around each presence and (3) 'target-group absences' (TGA), consisting of sites where other species of the group have been collected by the specialist, but not the species being modelled. To compare the predictive performance of the SDMs, the area under the curve statistic was calculated using an independent testing data set for each species.

Results MARS, MAXENT and LMR produce better results than the profile techniques. The models created with TGA are generally more accurate than those generated with pseudo-absences.

Main conclusions The advantages and disadvantages of different options for using pseudo-absences and TGA with profile and group discriminative modelling techniques are explained and recommendations are made for the future.

Keywords

Anthurium sp., group discriminative techniques, predictive performance, profile techniques, pseudo-absences, target-group absences.

*Correspondence: Rubén García Mateo, Universidad de Castilla-La Mancha, ICAM, Laboratorio de SIG y Teledetección, Av. Carlos III s/n, Toledo, 45071, Spain.
E-mail: Ruben.GMateo@uclm.es

INTRODUCTION

The distribution area of organisms is the fundamental basis for studies of biogeography, evolution, conservation, invasive species, design of protected areas, patterns of biodiversity or effects of climate change, among others (e.g., Thomas *et al.*, 2004; Thuiller *et al.*, 2005; Graham *et al.*, 2006; Jeschke & Strayer, 2008). However, species distributions are often poorly known, especially in tropical areas (Raven & Wilson, 1992). Ecological modelling has become a powerful tool that allows the generation of species distribution models (SDMs) to predict the likelihood of a given species occurring in areas for which data are either scarce or do not exist (Guisan & Zimmermann, 2000).

These SDMs are obtained through a series of methods that establish a relationship between different environmental variables and available data on the distribution of a given organism. In many cases, this distributional information is limited to that held by natural history collections (NHC) (Araújo & Williams, 2000; Barry & Elith, 2006). These collections record only locations where a species has been observed to be present, and there is no information available regarding the places where the species is not present (absences).

Some of the methods used in ecological modelling require absence or background data to generate SDMs (group discriminative techniques), whereas others are exclusively based on presence data (profile techniques). Although profile techniques use presence-only data, the lack of reliable absences also affects the performance of these methods. If a species is present in an insufficiently surveyed locality and remains undetected, the established 'profile' will be erroneous. The models generated with profile and group discriminative methods can be – and usually are – very different from each other, even when they are based on the same data (Loiselle *et al.*, 2003).

Although some data sets collected in the field include observed absences, it is quite difficult to verify that these are real absences (MacKenzie *et al.*, 2002; Graham *et al.*, 2004). Verifying the presence of a species at a site is feasible and objective, but verifying its absence is more complicated and subjective. In ecological modelling studies, absence data are useful when they reflect environmental conditions that do not allow for the development of the species being modelled (Lütolf *et al.*, 2006). However, absences because of dispersal limitations and/or historical factors should also be considered to avoid bias in the models generated (Jiménez-Valverde *et al.*, 2008). When considering the absence of a certain species in an area, the occurrence of errors may be as a result of several factors (Lütolf *et al.*, 2006), including: failing to detect the species in the site visited, which in turn can be a consequence of not having seen it or insufficient field work (time spent and/or area surveyed), identification errors, lack of observer experience, insufficient knowledge of the phenology or the biological cycle of the species concerned, or the effects of human activity (Phillips *et al.*, 2009). Real absences, when available, can reflect extinct or not yet established localities, in which case models will benefit by including variables capable of explaining such processes. Including these factors in the

modelling process is not easy although, and therefore some authors have taken a practical approach and considered absence data as a function that expresses zones where the species is less abundant (Brotons *et al.*, 2004) or even 'ambiguous data' (Rotenberry *et al.*, 2002) and untrue absence data.

There are few studies in which sampling has been designed specifically to obtain real absences with the objective of using them in ecological modelling studies (Elith, 2002; Elith *et al.*, 2006; T.P. Fera *et al.*, University of Texas-Pan America, Edinburg, unpublished data). This is costly in regard to both time and money, particularly in tropical areas. Therefore, to generate SDMs with group discriminative techniques, we must resort to unreal absences (Ponder *et al.*, 2001), generated at random within the study area and generally called 'pseudo-absences' (Zaniewski *et al.*, 2002). The manner in which these pseudo-absences are generated is very important, as they greatly influence the final model (Zaniewski *et al.*, 2002; Barry & Elith, 2006). Several ways of generating pseudo-absences have been proposed: (1) at random and establishing a buffer zone around the presences where pseudo-absences are not generated (Hirzel *et al.*, 2001), (2) in two steps, first generating an SDM with a group discriminative technique and random pseudo-absences, and then obtaining pseudo-absences only from the areas predicted to have higher suitability values (Zaniewski *et al.*, 2002), (3) also in two steps, but using a profile technique – Ecological Niche Factor Analysis (ENFA) – in the first step (Engler *et al.*, 2004) and (4) using distribution data from species – called 'auxiliary species' – with similar environmental requirements to the species being studied (Lütolf *et al.*, 2006). In all of these four strategies, the pseudo-absences may or may not be weighted to simulate a prevalence of 0.5 (i.e., Ferrier *et al.*, 2002).

Pseudo-absences are usually assumed to represent real absences, although they may include presence points (i.e. false absences), particularly when generated at random (Engler *et al.*, 2004). Consequently, pseudo-absences may represent biased or arbitrary data, and the SDMs so generated may not be reliable. Some studies have evaluated the performance of SDMs obtained with pseudo-absences (Hirzel *et al.*, 2001; Zaniewski *et al.*, 2002; Brotons *et al.*, 2004; Engler *et al.*, 2004; Pearce & Boyce, 2006; Chefaoui & Lobo, 2008). These studies have drawn contrasting conclusions, although most have recommended the use of group discriminative technique instead of profile techniques. Even in these studies, continuing research on the predictive performance of models created only with presence data is suggested (Pearce & Boyce, 2006).

In addition to the methods mentioned earlier, there are other ways to generate pseudo-absences that may be more effective than those generated at random. In this study, we focus on one such strategy, based on what we call 'target-group absences' (TGAs). To generate TGAs, we resort to the only information available on most occasions: NHC that represent presence-only data. These TGAs are similar to the 'inventory pseudo-absences' used in multiresponse methods, such as MARS-multiresponse (Elith *et al.*, 2006; Elith & Leathwick, 2007). These methods consider that if a site has been visited and a species has not been collected there, it can be considered an 'inventory

pseudo-absence. TGAs are localities where other species in the group of interest have been collected, but not the particular species being modelled. We hypothesize that these TGAs should be more discriminatory than pseudo-absences, and would result in more reliable models (Lütolf *et al.*, 2006). This approach has been scarcely used in ecological modelling studies (Lütolf *et al.*, 2006; Ferrier *et al.*, 2007; Phillips *et al.*, 2009).

The objectives of this study are therefore: (1) to develop a method for generating absences based on presence-only data housed in NHC, named TGAs, (2) to evaluate the predictive performance of models generated with TGAs, (3) to evaluate the predictive performance of models obtained using random pseudo-absences, (4) to compare both predictive performances to decide whether the use of random pseudo-absences should be rejected or not, (5) to compare the results obtained with six profile and group discriminative methods used in ecological modelling, to verify which techniques are the most accurate and (6) to evaluate the advantages and disadvantages of the different options.

METHODS

Biological data

The presence-only data used in this study is obtained from the TROPICOS database (Missouri Botanical Garden; <http://>

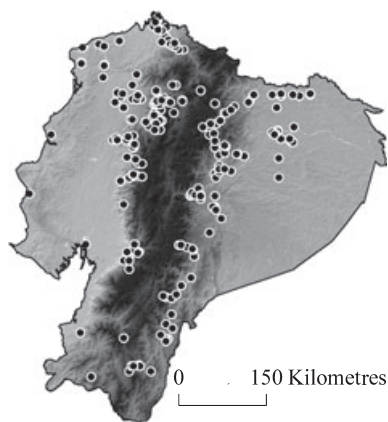


Figure 1 Thomas B. Croat's sampling localities (black points) in Ecuador included in TROPICOS.

mobot.mobot.org/W3T/Search/vast.html). We used all of the data available for the genus *Anthurium* (Araceae) in Ecuador. At the beginning of this study, we did not have documented absences for any of the species of the genus.

This genus has been thoroughly studied and collected in Ecuador by neotropical specialist, Thomas B. Croat (783 collections and 236 localities in TROPICOS; Croat, 1979, 1983, 1992, 1995, 1999) (Fig. 1). These collections were placed on a 0.0083° (~1 km) grid. In this transformation, some collections coincide on the same pixel, so that they represent a single presence (Table 1).

Among the Ecuadorian *Anthurium*, we selected species for which there were at least 18 presences collected by T. B. Croat. There were a total of five: *Anthurium dolichostachyum* Sodirol, *A. harlingianum* Croat, *A. propinquum* Sodirol, *Anthurium truncicola* Engl. and *A. versicolor* Sodirol. The five species show different geographical distributions, and consequently also different ecological requirements: (1) *A. dolichostachyum* (0–2000 m) is endemic to the coast and western side of the Andes, (2) *A. harlingianum* (0–2000 m) grows in the Amazonian and at the foot of the Eastern Andes, (3) *A. propinquum* (0–2000 m) grows on the coast and the foot of the Western Andes, (4) *A. truncicola* (0–2500 m) grows at the foot of both slopes of the Andes and less frequently, on the coast and the Amazon basin and finally, (5) *A. versicolor* (0–2500 m) that grows at the foot of the Andes, both on the eastern and western slopes, is frequently found on the northern coast (Esmeraldas Province) and sporadically in the Amazonian (Croat, 1999).

Generation of pseudo-absences and target-group absences

For the training data set (Fig. 2), specimens collected by T. B. Croat ('presences croat', PC) were used as presences, and absences were generated either as TGA (generated and selected at random) or as pseudo-absences generated at random, either in equal number to the available presences ('unweighted' data set) or in a much larger number and weighted to simulate 0.5 prevalence ('weighted' data set). We used this intermediate level of prevalence to avoid statistical artefacts because of differences in sampling prevalence, known to produce results biased to the larger group in logistic regression (Hosmer & Lemeshow, 2000). Several authors have shown that interme-

Table 1 Presence and absence data for each species included in this study. C, collections in TROPICOS; P, unique presences at 0.0083° spatial resolution; PC, unique presences by T. B. Croat (training data sets); PSA, unweighted random pseudo-absences (training data sets); PSB, weighted random pseudo-absences (training data sets); TGA, target-group absences (training data sets); P-V, presences by other authors (testing data sets); TGA-V, target-group absences (testing data sets).

Species	C	P	PC	PSA	TGA	P-V	TGA-V	PSB
<i>Anthurium dolichostachyum</i>	106	79	47	37	42	32	20	7422
<i>A. harlingianum</i>	39	32	18	19	15	14	22	7876
<i>A. propinquum</i>	39	34	22	26	22	12	21	8125
<i>A. truncicola</i>	66	52	23	29	23	29	18	7311
<i>A. versicolor</i>	126	97	44	37	43	53	29	6454

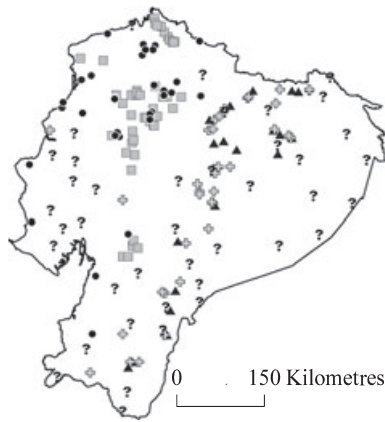


Figure 2 Presence and absence data for *Anthurium dolichostachyum*. Grey squares = presences by T. B. Croat (training data set); black question marks = unweighted pseudo-absences (training data set); grey crosses = target-group absences (training data set); black points = presences by other authors (testing data set); dark grey triangles = target-group absences (testing data set). The 7422 randomly generated weighted pseudo-absences used to build the models are not depicted on this map.

diate prevalence provides better models (Fielding & Bell, 1997; Cumming, 2000; Olden *et al.*, 2002; McPherson *et al.*, 2004).

Target-group absences were localities in TROPICOS where T. B. Croat collected *Anthurium* but did not find the species being modelled. It is assumed that a specialist would collect all species of his/her group at the site being visited, and sites where the species is not collected represent areas where the probability of finding such species is very low and consequently represent 'absences' in the sense of Brotons *et al.* (2004) (Table 1). In fact, when field work is carried out with the objective of collecting absence data, the procedure is similar (T.P. Feria, B. Loiselle *et al.*, unpublished data).

Unweighted random pseudo-absences (PSA) were generated in ArcView 3.2 with the extension 'Random Point Generator 1.28' with the following constraints: (1) we generated approximately the same number of pseudo-absences as presences to avoid problems associated with unbalanced prevalence (Titeux, 2006). (2) To avoid spatial autocorrelation issues and also to collect information on the different ecological conditions in the study area, we defined a minimum distance of 30 km between pseudo-absences (Pearson *et al.*, 2006; Elith & Leathwick, 2007). (3) To avoid increasing the false-negative rate, we defined a buffer of 30 km in diameter around each presence from where pseudo-absences were eliminated (Fig. 3; Anderson, 2003; Loiselle *et al.*, 2003), except in the case of MAXENT, because the technique is designed to work with presence versus background data, and therefore assumes that the background data will include both absences and presences (Phillips *et al.*, 2009).

The size of the buffer could be a random measurement, or it could be established according to a particular characteristic of the species, such as the dispersal capacity (Graham & Hijmans, 2006). In our case, the distance of 30 km was calculated based on the information contained in the maps according to the pixel

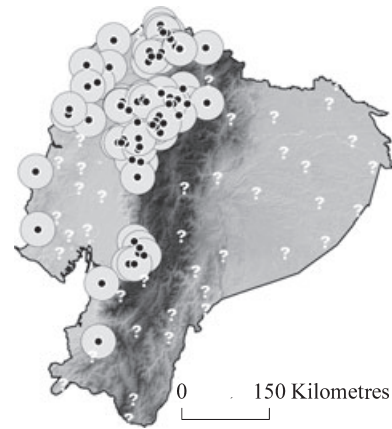


Figure 3 Presences of *Anthurium dolichostachyum* (black points), 30 km buffer area (grey circles) and unweighted pseudo-absences (question marks) generated for this species.

size. The pixel size of a set of SDMs was duplicated consecutively, and the information contained in each of the rescaled maps was calculated using Shannon's formula for entropy:

$$H = \sum_{i=1}^n p(i) \cdot \log p(i).$$

Our results (Fig. 4) show that the information contained in the rescaled maps remains approximately constant until it reaches a pixel size of 0.256° (~ 32 km at the Equator), and this value was chosen as the radius of the buffer.

Weighted random pseudo-absences (PSB) were generated for the unweighted pseudo-absences mentioned earlier, except that we generated 10,000, from which those inside a 30-km buffer around each presence were eliminated (Table 1). In all subsequent analyses, they were weighted to simulate a prevalence of 0.5.

Environmental predictors

The environmental predictors used as independent variables in the models were the 19 bioclimatic variables (Hijmans *et al.*,

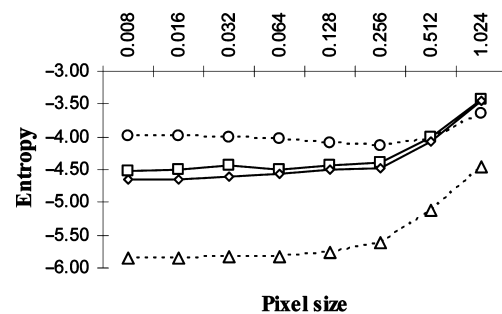


Figure 4 Relationship between the information contained in a habitat suitability map and the cell size for four species. The information in the models remains constant until a pixel size of 0.256° (~ 32 km) is reached, when there is a sharp decrease in the contained information.

2005) from Worldclim 1.3 (<http://www.worldclim.org>), derived from average monthly temperature and rainfall surfaces linked to a global digital elevation model.

Ecological modelling methods

The SDMs were generated with six different methods, all of them well known; Elith *et al.* (2006) provide good description, references, parameterization and software for using these methods. The only exception is that we used Salford Systems' MARS 2.0 (<http://www.salford-systems.com>) software instead of the 'mda' R package, as we have tested this software before (Muñoz & Felicísimo, 2004).

These methods can be classified into three main types:

Group discriminative techniques

In this study, we have used three group discriminative methods: logistic multiple regression (LMR) and multivariate adaptive regression splines (MARS, Friedman, 1991) – both of which require absences – and MAXENT (i.e. maximum entropy, Phillips *et al.*, 2006; Phillips & Dudík, 2008), which works with background points. The first has a longer history in ecological modelling (e.g., Pereira & Itami, 1991). MAXENT and MARS are more recent additions to the modeller's toolbox, both exhibiting good performance (Muñoz & Felicísimo, 2004; Elith *et al.*, 2006; Leathwick *et al.*, 2006). For group discriminative techniques, we generated SDMs using three types of absences: unweighted pseudo-absences in equal number to presences, weighted pseudo-absences in much larger number than presences and TGA. Whereas LMR and MARS need absence data generated by the analyst, MAXENT can either work with internally generated absences, called 'background points', or be forced to work with absences generated by the user. To have a comparable data set across methods, we forced MAXENT to create the models with the same absences as LMR and MARS through 'swd' files (cf. <http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc> or MAXENT software help). However, MAXENT is designed to work with presence versus background data, rather than presence versus absence data (Phillips *et al.*, 2009). Therefore, unlike the absence samples generated for the two regression techniques (LMR and MARS), presences were not excluded from the samples generated for use with MAXENT. In the case of pseudo-absences, this meant we ignored the 3rd constraint described above (i.e. defining a 30-km buffer around presences). For the target-group strategy, we included all localities where any species of *Anthurium* was collected, including the particular species being modelled.

Profile techniques

These techniques use only the information available on presences to generate the SDMs. We used two methods: an environmental envelope technique, BIOCLIM (Busby, 1986, 1991), and the Gower's point-to-point similarity metric

(Carpenter *et al.*, 1993). These are the two techniques that are often used in ecological modelling.

Mixed technique

Genetic algorithm for rule-set production (Stockwell & Peters, 1999), widely used in ecological modelling, generates a set of rules (atomics, regression logic, environmental envelopes, range rules, etc.) to build the SDM, some of which use only presence data, whereas others also need absences. Genetic algorithm for rule-set production (GARP) self-generates its own pseudo-absences ('background points'). Unlike MAXENT, there is no way to force GARP to use purposefully generated background points.

Model evaluation

Once all the SDMs were generated (12 for each species), we measured their performance using the testing data set (Fig. 2) to calculate the area under the curve (AUC) statistic (Hanley & McNeil, 1982). To create this data set, we used as presences the sites where collectors other than T. B. Croat had collected the species being modelled (Table 1, P-V), and as absences randomly selected TGA obtained as mentioned above and not used in the training data set (Table 1, TGA-V). To avoid bias in specimen identification, all the collections by other authors were reviewed by T. B. Croat (Croat, 1999).

RESULTS

Comparison of modelling techniques

In Fig. 5 we present, as an example, SDMs obtained using the six different modelling techniques for *A. dolichostachyum*, the species with the most presences. Large incongruities between techniques are evident. While not presented in detail in this study, the same incongruities were also evident for the remaining species in this work. AUC values were generally higher for group discriminative techniques (LMR, MARS and MAXENT) than that for profile (BIOCLIM, Gower's distance) or the mixed technique GARP (Table 2). Nine of the best ten models (two for species; Table 2, in bold) were generated with group discriminative techniques (LMR, MARS or MAXENT).

Comparison of pseudo-absence and target-group absence strategies

A key result of this study is that four of the five best models (Table 2, underlined) were obtained using TGA, and only one using pseudo-absences in equal number to the presences. Although the sample size is small, according to a binomial test (TGA, $k = 15$; non-TGA, $k' = 30$), the probability of obtaining, through chance alone, at least four best models using TGA is $P = 0.0453$. This value is just less than the conventional critical 0.05 level, suggesting that the values in Table 2 are not random. Among the group discriminative techniques, MAXENT

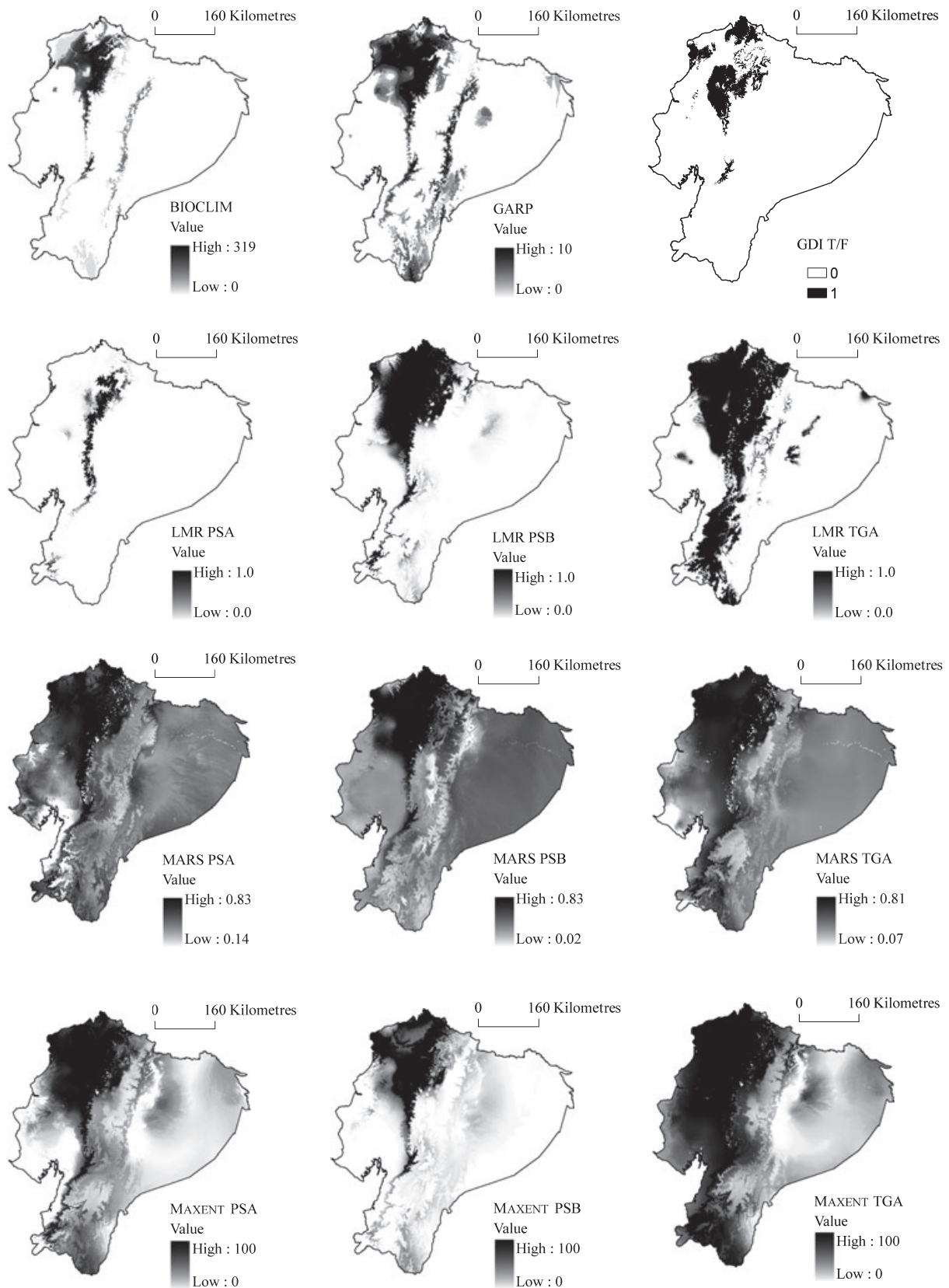


Figure 5 Habitat suitability maps obtained for *Anthurium dolichostachyum* using 12 different combinations of modelling method (BIOCLIM, GARP, Gower's distance index, logistic multiple regression and MAXENT) and type of absence data. GDI T/F = Gower's distance index true/false (presence/absence); PSA = unweighted pseudo-absences; PSB = weighted pseudo-absences; TGA = target-group absences.

Table 2 AUC values of the 12 SDMs generated for each of the five species. The AUC statistics were calculated using a testing data set. We ran these models using different strategies for generating absence data and different modelling methods. Absence data used: target-group absences (TGA); unweighted pseudo-absences in approximately the same number as presences (PSA); approximately 10,000 (see Table 1) pseudo-absences (PSB). Methods used: BIOCLIM; Gower's distance index (GDI); genetic algorithm for rule-set production (GARP); MAXENT; multivariate adaptive regression splines (MARS); logistic multiple regression (LMR). The two best values for each species are in bold, the highest underlined.

	BIOCLIM	GDI	GARP	MAXENT			MARS			LMR		
				TGA	PSA	PSB	TGA	PSA	PSB	TGA	PSA	PSB
<i>Anthurium dolichostachyum</i>	0.770	0.804	0.816	<u>0.977</u>	0.791	0.897	0.909	0.873	0.823	0.961	0.853	0.841
<i>A. harlingianum</i>	0.786	0.932	0.813	<u>0.912</u>	0.904	0.860	<u>0.968</u>	0.877	0.867	0.929	0.864	0.909
<i>A. propinquum</i>	0.750	0.800	0.810	<u>1.000</u>	0.988	0.994	<u>0.972</u>	0.984	0.988	0.851	0.873	0.992
<i>A. truncicola</i>	0.708	0.759	0.602	<u>0.801</u>	0.718	0.722	0.667	0.655	0.519	0.812	<u>0.828</u>	0.739
<i>A. versicolor</i>	0.570	0.694	0.609	0.664	0.739	0.651	<u>0.862</u>	0.635	0.673	0.692	<u>0.772</u>	0.763

and MARS both performed better using TGA than using pseudo-absences for four of the five species. This difference was less apparent for LMR. No clear trend was discernable in the relative performance of the two pseudo-absence strategies for any of the three group discriminative techniques.

DISCUSSION

Comparison of modelling techniques

Higher AUC values were obtained with methods that use both presences and absences (LMR, MARS and MAXENT), a phenomenon that has been already documented by other authors (Guisan *et al.*, 2002; Brotons *et al.*, 2004; Segurado & Araújo, 2004). In these methods, the absences help to define the areas where the species does not thrive (Brotons *et al.*, 2004). Use of these absences (either pseudo-absences or TGA) will result in improved models if they correctly represent areas that are unfavourable to the development of the species (Brotons *et al.*, 2004), but may result in increased error and reduced predictive power if they are false absences (Solow, 1993; Welsh *et al.*, 1996). The methods with lower AUC values were GARP, Gower similarity and BIOCLIM, which agrees with the results obtained by Elith *et al.* (2006), who showed that more recently developed algorithms, such as MARS or MAXENT, give better results.

If we analyse the suitability map generated for *A. dolichostachyum* (Fig. 5), and we take into account that for this species the algorithms that offer the highest AUC values are LMR, MARS and MAXENT, it becomes clear that the lower AUC values obtained with Gower's metric and BIOCLIM are a consequence of overfitting (Engler *et al.*, 2004; Chefaoui & Lobo, 2008), resulting in an increased level of omission error. However, with GARP the opposite situation occurs, as this technique tends to overpredict the distribution of the species, resulting in an increased level of commission error (Hernandez *et al.*, 2006). In the case of Gower's metric and BIOCLIM, this effect could be a consequence of not using absences and, as a result, not using climatic information about unfavourable sites (Pearson *et al.*, 2006; Chefaoui & Lobo, 2008; Phillips *et al.*, 2009).

The Gower's metric is the profile method reaching AUC values closest to the discriminative techniques (Table 2), and even sometimes reaches higher AUC values, e.g. relative to SDMs generated with pseudo-absences and MARS for *A. harlingianum*, *A. truncicola* and *A. versicolor*, or LMR for *A. harlingianum*. Jiménez-Valverde *et al.* (2008) state that profile methods generally predict distributions closer to the potential, rather than realized distribution of a species. The reduction in accuracy in our study may, therefore, be as a result of the fact that our testing data sets include data only on the species' realized distribution.

The appropriateness of different modelling methods depends on the aim of any given study. MARS has the advantage that fitted models are relatively easy to interpret, both from a statistical and an ecological point of view (Muñoz & Felicísimo, 2004; Austin, 2007; Elith & Leathwick, 2007). MAXENT, on the contrary, can automatically generate its own pseudo-absences and seems to produce stable models with very good predictive performance for small sample sizes (Hernandez *et al.*, 2006; Phillips *et al.*, 2006; Papeş & Gaubert, 2007; Pearson *et al.*, 2007). We conclude that MARS is the most appropriate when we wish to perform a detailed analysis and we have data for generating TGA, whereas MAXENT is recommended when we want to model a large number of species, we do not need a thorough analysis or for exploratory modelling when only few presences are available (Pearson *et al.*, 2007; Wisz *et al.*, 2008).

Comparison of pseudo-absence and target-group absence strategies

Species distribution models generated with either pseudo-absences or TGA had generally high AUC values – 37 of the 45 SDMs generated had an AUC over 0.7, a value considered in the literature as sufficient for using the generated SDMs for conservation purposes (Pearce & Ferrier, 2000; Elith & Leathwick, 2007), although other authors disagree (Lobo *et al.*, 2008; Peterson *et al.*, 2008).

Our results confirm that SDMs generated with TGA, in general, have greater predictive performance than SDMs

generated with purely random pseudo-absences. Similar results have been obtained by Phillips *et al.* (2009), who generated 'target-group background' points from '...the presence localities for all species in the same target-group.' These authors consider that such localities reflect the same bias as the presence-only data and, therefore, improve the performance of the models. On the contrary, other authors have claimed that the environmental bias in collections does not greatly affect distribution predictions (Kadmon *et al.*, 2004; Loiselle *et al.*, 2008).

We have in this study chosen absences through a criterion that requires knowledge of collection effort, because our objective is to generate reliable absences. On the contrary, Phillips *et al.* (2009) chose absences from localities where the species was not collected, but without considering collection effort, as their aim was different: to equalize the bias for absences and presences. Bias may influence the reliability of the generated models, but in this study, we focus on the importance of generating reliable absences as a key strategy for producing high-quality models (Lobo, 2008).

The better results obtained with TGA compared with that of pseudo-absences could be as a result of a combination of various factors: (1) TGA provide more accurate data on environments where a species does not occur, than do pseudo-absences generated at random, and therefore result in a lower proportion of false negatives (i.e. higher specificity). (2) TGA represent the same bias as the presence-only data, which eliminate inaccuracies in resulting models because of differences in bias between occurrence collection and pseudo-absences or background points (Phillips *et al.*, 2009). (3) The collection localities used as absences in the testing data set are geographically closer to the TGA used to build the model (most collectors tend to re-visit already collected areas) than they are to pseudo-absences (Fig. 2). In other words, TGA used to generate and validate the model are not independent, which may inflate resulting AUC values to some unknown extent (Phillips *et al.*, 2009).

Unfortunately, in many situations insufficient data will be available to enable the generation of TGA. In such situations, pseudo-absences generated at random, excluding a buffer area around each presence, seem to be a good solution. Pseudo-absences have two additional advantages: (1) the entire area of the study can be sampled, adding useful information to build the model, and (2) they have a high probability of being correctly located in the case of an environmental restricted species. This may be the reason why the SDM generated with pseudo-absences present better results than the SDM generated with TGA in some cases (Elith & Leathwick, 2007; Chéfaoui & Lobo, 2008). We conclude that it is preferable to use pseudo-absences if the number of TGA that can be generated is small.

In this study, we have used the genus *Anthurium* as a case study. Data sets as complete as this are not common, but TGA can be generated using simpler or more common sets of data housed in NHC by: (1) using data from broader groups (Phillips *et al.*, 2009), e.g. all vascular plants rather than a single genus, or (2) considering sites that have been visited by any collector and where the target species has not been

collected, similar to the 'inventory pseudo-absence' of multi-response methods (Elith & Leathwick, 2007).

Comparing results between species

Species distribution models for species with wide ranges and ecological tolerances usually have less predictive power than those for species with more restricted distribution ranges and/or ecological tolerances (Manel *et al.*, 2001; McPherson *et al.*, 2004; Luoto *et al.*, 2005; Elith *et al.*, 2006; T.P. Ferial, B. Loiselle *et al.*, unpublished data). This is confirmed by our results. Models for the two species with the widest distribution (*A. truncicola* and *A. versicolor*, growing on both slopes of the Andes) performed poorly, regardless of the method employed, relative to the other three taxa with more restricted ranges, although these three species have relatively small sample sizes. This confirms the previous observation by Hernandez *et al.* (2006) that the performance of SDMs is, in most cases, determined more by the ecology of the species than by sample size.

CONCLUSIONS

From our results, we conclude that when working with data sets derived from NHC, SDMs generated with group discriminative techniques (LMR, MARS and MAXENT) are more reliable than models generated with profile or mixed techniques (BIOCLIM, Gower's similarity metric and GARP). For the group discriminative techniques, models generated with TGA generally perform better than models derived using pseudo-absences. Each option considered presents advantages and drawbacks, and its use for a particular study depends on the objective and on data availability. When enough data are available, it seems appropriate to: (1) use group discriminative techniques as they are more reliable than profile techniques, (2) use MARS or other regression techniques if the aim is to perform an in-depth analysis of each species, or MAXENT when a less intensive analysis of large numbers of species is required, (3) use TGA instead of pseudo-absences if sufficient data are available to generate them and (4) if forced to use pseudo-absences, create a buffer around each presence to minimize the false-negative rate. We also conclude that: (1) absence data help to define areas of low suitability for the species of interest, (2) Gower's similarity metric and BIOCLIM seem to overfit the training data, although Gower's similarity metric is the profile technique that obtains the better results and is perhaps a method of choice if there are very few presences and (3) GARP seems to have a tendency towards overprediction.

ACKNOWLEDGEMENTS

We thank the BBVA Foundation for its financial support and the Missouri Botanical Garden for generously providing the *Anthurium* data. We thank three anonymous referees and especially the Editor, S. Ferrier, for their comments, which allowed us to improve the manuscript greatly.

REFERENCES

- Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: test for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30**, 591–605.
- Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.
- Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.
- Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.
- Brottons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.
- Busby, J.R. (1986) *Bioclimate Prediction System (BIOCLIM). User's Manual Version 2.0*. Australian Biological Resources Study Leaflet, Canberra, Australia.
- Busby, J.R. (1991). BIOCLIM: a bioclimate analysis and prediction system. *Nature conservation: cost effective biological surveys and data analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Melbourne, Australia.
- Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.
- Chefaoui, R.M. & Lobo, J.M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.
- Croat, T.B. (1979). The distribution of the Araceae. *Tropical botany* (ed. by K. Larsen and L.B. Holm-Nielsen), pp. 291–308. Academic Press, London.
- Croat, T.B. (1983) A revision of the genus *Anthurium* (Araceae) of Mexico and Central America. Part I: Mexico and Middle America. *Annals of the Missouri Botanical Garden*, **70**, 211–420.
- Croat, T.B. (1992) Species diversity of Araceae in Colombia: a preliminary survey. *Annals of the Missouri Botanical Garden*, **79**, 17–28.
- Croat, T.B. (1995). Floristic comparisons of Araceae in six Ecuadorian florulas. *Biodiversity and conservation of Neotropical Montane Forest* (ed. by S.P. Churchill, H. Balslev, E. Forero and J.L. Luteyn), pp. 489–499. The New York Botanical Garden, New York.
- Croat, T.B. (1999). Araceae. *Catalogue of the vascular plants of Ecuador* (ed. by P.M. Jørgensen and S. León-Yáñez), pp. 227–246. Missouri Botanical Garden, St. Louis.
- Cumming, G.S. (2000) Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*, **27**, 425–440.
- Elith, J. (2002). *Predicting the distribution of plants*. PhD Dissertation, School of Botany, The University of Melbourne, Melbourne, Australia.
- Elith, J. & Leathwick, J.R. (2007) Predicting species distributions from museum and herbarium records using multi-response models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, **13**, 265–275.
- Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
- Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.
- Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.
- Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalised dissimilarity modelling to analyse and predict patterns of beta-diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264.
- Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.
- Friedman, J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**, 1–141.
- Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.
- Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Perteon, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Ecology and Evolution*, **19**, 497–503.
- Graham, C.H., Moritz, C. & Williams, S.E. (2006) Habitat history improves prediction of biodiversity in rainforest fauna. *Proceedings of the National Academy of Sciences USA*, **103**, 632–636.
- Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.
- Guisan, A., Edwards, T.C., Jr & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.
- Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.
- Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.
- Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.
- Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.
- Hosmer, D.W. & Lemeshow, S. (2000). *Applied logistic regression*, 2nd edn. John Wiley & Sons, New York.

- Jeschke, J.M. & Strayer, D.L. (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences*, **1134**, 1–24.
- Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, **14**, 885–890.
- Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.
- Leathwick, J.R., Elith, J. & Hastie, T. (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, **199**, 188–196.
- Lobo, J.M. (2008) More complex distribution models or more representative data? *Biodiversity Informatics*, **5**, 14–19.
- Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.
- Loiselle, B., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams, P.H. (2003) Avoiding pitfalls of using species distributions models in conservation planning. *Conservation Biology*, **17**, 1591–1600.
- Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.
- Luoto, M., Heikkinen, R.K. & Saarinen, K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575–584.
- Lütolf, M., Kienast, F. & Guisan, A. (2006) The ghost of past species occurrence: improving species distributions models for presence-only data. *Journal of Applied Ecology*, **43**, 802–815.
- MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.
- Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence-absence models in ecology: the need to account prevalence. *Journal of Applied Ecology*, **38**, 921–931.
- McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? *Journal of Applied Ecology*, **41**, 811–823.
- Muñoz, J. & Felicísimo, A.M. (2004) A comparison between some statistical methods commonly used in predictive modeling. *Journal of Vegetation Science*, **15**, 285–292.
- Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.
- Papeş, M. & Gaubert, P. (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions*, **13**, 890–902.
- Pearce, J. & Boyce, M. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.
- Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.
- Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C.J., Miles, L., Segurado, P., Dawson, T.P. & Lees, D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704–1711.
- Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.
- Pereira, J.M.C. & Itami, R.M. (1991) GIS-based habitat modelling using logistic multiple regression: a study of the Mt. Graham red squirrel. *Photogrammetric engineering & Remote sensing*, **57**, 1475–1486.
- Peterson, A.T., Papes, M. & Soberón, J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, **213**, 63–72.
- Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.
- Phillips, S.J., Anderson, R.P. & Schapire, R.P. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.
- Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only models of species distributions: implications for selection of background and pseudo-absences. *Ecological Applications*, **19**, 181–197.
- Ponder, W.F., Carter, G.A., Flemons, P. & Chapman, R.R. (2001) Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, **15**, 648–657.
- Raven, P.H. & Wilson, E. (1992) A fifty-year plan for biodiversity surveys. *Science*, **258**, 1099–1100.
- Rotenberry, J.T., Knick, S.T. & Dunn, J.E. (2002) A minimalist approach to mapping species' habitat: pearson's planes of closest fit. *Predicting species occurrences issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund and M.L. Morrison), pp. 218–289. Island Press, Washington.
- Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.
- Solow, A.R. (1993) Inferring extinction from sighting data. *Ecology*, **74**, 962–964.
- Stockwell, D. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, **13**, 143–158.
- Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y., Erasmus, B.F.N., de Sique-

- ira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L.J., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O. & Williams, S.E. (2004) Extinction risk from climate change. *Nature*, **427**, 145–148.
- Thuiller, W., Lavorel, S., Araujo, M.B., Sykes, M.T. & Prentice, I.C. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.
- Titeux, N. (2006). *Modelling species distribution when habitat occupancy departs from suitability. Application to birds in a landscape context*. Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Welsh, A.H., Cunningham, R.B., Donnelly, C.F. & Lindenmayer, D.B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.
- Wisn, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & Group, N.P.S.D.W. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

- Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

BIOSKETCH

Dr. Rubén G. Mateo: His main scientific activity is ecological modelling applied to the study of biodiversity, biogeography, conservation biology and fire ecology. He is currently working on the optimization of different methods and techniques to produce reliable species distribution models. He is also involved in floristic and fire ecology studies of central Iberian Peninsula.

Author contributions: R.G.M. conceived the ideas; T.B.C. collected the data; R.G.M., A.M.F. and J.M. analysed the data; and R.G.M. and J.M. led the writing.

Editor: Simon Ferrier